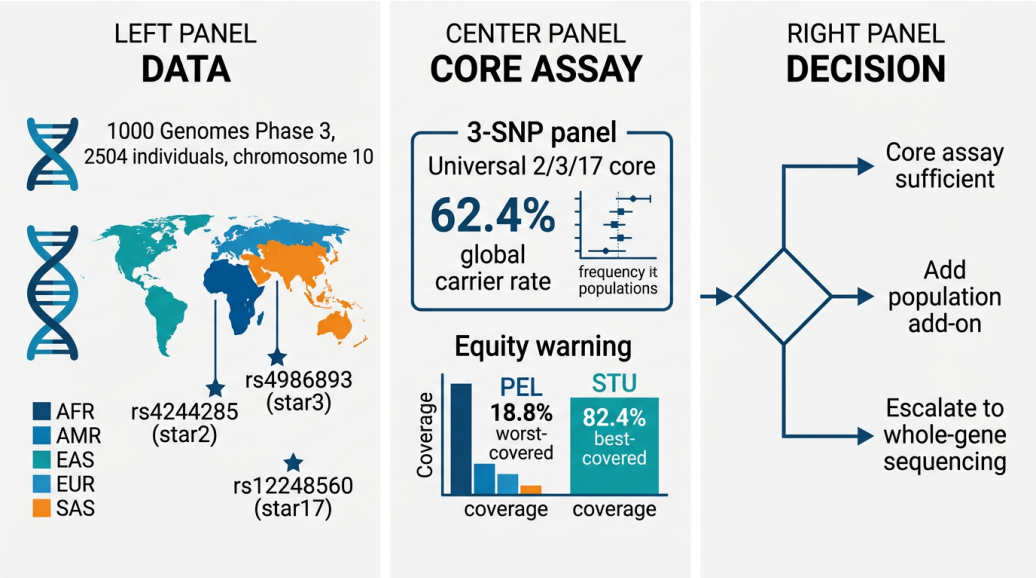
Designing an Equitable, Low-Cost CYP2C19 Genotyping Assay

Universal *2/*3/*17 Core versus Population-Specific Add-ons: A Population-Genetic Coverage and Equity Audit

A Research-Only Technical Report

Reference cohort: 1000 Genomes Project Phase 3 (release 20130502), GRCh37
2,504 individuals · 5 super-populations · 26 sub-populations · chromosome 10

Variant annotation: dbSNP Build 157 · Guideline reference: ClinPGx/CPIC record PA166251443



Graphical abstract. The universal three-SNP core assay, its global coverage, and the evidence-based decision framework governing escalation to population-specific add-ons or whole-gene sequencing.

Scope and non-clinical disclaimer. This document is a research-only assay-design and population-genetics analysis. It reports allele frequencies, coverage, statistical uncertainty, and an equity audit. **It does not provide, and must not be used as, prescribing, dosing, or diagnostic advice for any individual patient.** All clinical decisions remain the responsibility of qualified clinicians acting under validated, regulatory-approved assays and current guidelines.

Generated 2026-07-12 · Analysis provenance: verified six-step computational workflow (Steps 1–6)

Executive Summary

Decision. For a low-cost, research-only CYP2C19 genotyping assay, we recommend deploying a **universal three-SNP core panel** capturing *2 (rs4244285), *3 (rs4986893), and *17 (rs12248560) as the first-line design, **augmented by a defined, trigger-activated add-on module** (*4, *6, *8) rather than a single fixed universal panel. The core is cheap, technically near-perfect, and globally efficient, but its clinical yield is profoundly ancestry-dependent, and a rigid universal-only design would encode a measurable equity gap.

This report interrogates a foundational trade-off in pharmacogenomic assay design: whether a minimal, inexpensive genotyping panel built from the three canonical CYP2C19 tag SNPs is sufficient across globally diverse populations, or whether it must be supplemented with population-specific content to remain equitable. Using the phased genotypes of all 2,504 individuals in the 1000 Genomes Project Phase 3 reference panel, we estimated phased allele and haplotype frequencies for every super- and sub-population, audited genotype missingness and Hardy–Weinberg equilibrium (HWE), quantified how faithfully each SNP tags its designated star allele, and compared total (global) coverage against worst-population coverage.

The core panel performs superbly on technical grounds. Genotype missingness was zero, no locus deviated from HWE in any of 32 population strata after false-discovery-rate correction, and each SNP tagged its star allele with sensitivity of essentially 1.0 and positive predictive value ≥ 0.999. Globally, the three-SNP core identifies at least one non-*1 star allele in **62.4%** of individuals (95% Wilson confidence interval [CI]: 60.5–64.3%). Yet this global average conceals a threefold spread across populations: carrier rates range from **82.4%** in the Sri Lankan Tamil (STU) sub-population down to just **18.8%** (95% CI: 11.9–28.4%) in the Peruvian (PEL) sub-population—an absolute gap of 43.60 percentage points. A separate equity audit of loss-of-function alleles *not* covered by the core (*4, *8) identified eight individuals globally who would be misclassified as normal metabolizers, and revealed a structural blind spot: the canonical *6 variant is absent from the Phase 3 call set entirely, so array- and imputation-based reference data cannot even measure it.

These findings motivate a decision framework, not a single answer. The core is the correct universal baseline; ancestry, study-endpoint, and technology triggers determine when the *4/*6/*8 add-on or a transition to whole-gene sequencing becomes warranted. We provide the assay specification, per-population coverage with uncertainty, the equity audit, and an explicit statement of the evidence that would change the design.

Contents

1 Introduction and Background

1.1 CYP2C19 as a pharmacogenomic exemplar

The cytochrome P450 2C19 enzyme (CYP2C19) metabolizes a clinically important set of substrates, and the gene encoding it is among the most actively curated pharmacogenes in translational medicine [1, 2]. Its allelic architecture is organized under the star (*) nomenclature maintained by the Pharmacogene Variation (PharmVar) Consortium, in which haplotypes defined by one or more variants are assigned functional labels ranging from no function through normal to increased function [3]. The Clinical Pharmacogenetics Implementation Consortium (CPIC) translates these genotype-derived phenotypes into evidence-graded therapeutic guidance for several drug classes, including antiplatelet therapy, proton-pump inhibitors, and selected antidepressants [4–7]. The CPIC/ClinPGx guideline record for CYP2C19 and clopidogrel, `PA166251443`, anchors the functional interpretation used throughout this report [4, 8].

Three star alleles dominate the clinically actionable variation. The `*2` allele, tagged by `rs4244285` (a synonymous change that creates a cryptic splice site abolishing enzyme activity), is the most common no-function allele worldwide and was the original molecular explanation for the poor-metabolizer phenotype of S-mephenytoin [9]. The `*3` allele, tagged by `rs4986893` (a premature stop codon), is a second no-function allele largely restricted to East Asian ancestry. The `*17` allele, tagged by the promoter variant `rs12248560`, increases transcription and confers an increased-function (rapid/ultrarapid metabolizer) phenotype [10]. Because these three variants sit on mutually near-exclusive haplotype backgrounds, a compact panel of three tag SNPs can, in principle, reconstruct the majority of clinically relevant diplotypes at very low cost.

1.2 The equity problem in minimal pharmacogenomic panels

Minimal genotyping panels are attractive for research and resource-limited settings because they are inexpensive, robust, and easy to multiplex. However, they encode a prior assumption: that the handful of variants selected—overwhelmingly discovered in European and East Asian cohorts—adequately represent the actionable variation in *all* populations. A large body of work has documented that human genetics research remains heavily Eurocentric, and that this imbalance propagates directly into clinical tools, from polygenic scores to variant-pathogenicity calls and pharmacogenomic panels [11–13]. For CYP2C19 specifically, worldwide surveys show substantial inter-ethnic variation in star-allele frequencies and in the rare, sometimes population-private, loss-of-function alleles that a fixed panel will systematically miss [14–16]. African populations in particular carry no-function alleles—including `*6`, `*8`, and others—that are invisible to a `*2/*3/*17` core, and continental capacity-building efforts such as H3Africa exist precisely because reference data for these populations remain sparse [17].

1.3 Objectives

This report addresses a single, concrete design question with population-genetic evidence:

> *Should a low-cost, research-only CYP2C19 assay use a universal `*2/*3/*17` core, or does equitable performance require population-specific add-ons?*

To answer it we (i) estimate phased allele and haplotype frequencies by population; (ii) audit genotype missingness and HWE deviations as quality-control checks; (iii) test whether each SNP faithfully tags its stated star allele; (iv) compare total versus worst-population coverage with formal uncertainty; and (v) conduct an equity audit of the loss-of-function variation the core cannot see. The deliverables are an assay specification, per-population coverage with uncertainty, an equity audit, and a decision framework enumerating the evidence that would change the design. We reiterate that the analysis is descriptive and design-oriented; it provides no prescribing advice.

2 Materials and Methods

2.1 Reference data and cohort

Phased genotype data were obtained from the 1000 Genomes Project Phase 3 integrated call set (release 20130502, GRCh37 assembly), specifically the chromosome-10 variant file `ALL.chr10.phase3_shapeit2_mvncall_integrated_v5b` together with the `integrated_call_samples_v3.20130502.ALL.panel` sample-annotation file [18, 19]. The call set comprises 2,504 unrelated individuals (5,008 phased chromosomes) drawn from five super-populations—African (AFR, $n = 661$), admixed American (AMR, $n = 347$), East Asian (EAS, $n = 504$), European (EUR, $n = 503$), and South Asian (SAS, $n = 489$)—and 26 sub-populations. Haplotypes in this release were statistically phased with SHAPEIT2, permitting direct cis/trans resolution of the tag SNPs [20]. Genotype records were parsed using variant-call-format tooling and cross-checked against expected sample counts [21].

2.2 Variant identification and coordinate verification

Each target SNP was verified against dbSNP Build 157 to confirm its authoritative GRCh37 and GRCh38 coordinates, reference/alternate alleles, and sequence-identifier mapping before extraction [22]. This step proved essential: the task-provided GRCh37 coordinates for all three core SNPs did *not* match dbSNP and would have led to extraction of the wrong loci. We therefore used the dbSNP-authoritative positions throughout: `rs4244285` at chr10:96,541,616 (G>A, `*2`), `rs4986893` at chr10:96,540,410 (G>A, `*3`), and `rs12248560` at chr10:96,521,657 (C>T, `*17`). The GRCh37→GRCh38 liftover offset was internally consistent across all three core SNPs (a constant 1 759 757.00 bp), providing an additional coordinate-integrity check. Add-on-allele identifiers were validated identically (Section 2.7).

2.3 Allele, haplotype, and diplotype frequency estimation

Because the release is physically phased, star-allele haplotypes were read directly from the phased chromosomes rather than inferred by an expectation–maximization algorithm. Each chromosome was encoded as a three-bit vector over {`rs4244285`, `rs4986893`, `rs12248560`}, and mapped to a star allele under the definitions in Table 1: `000`→`*1` (normal function), `100`→`*2`, `010`→`*3`, `001`→`*17`, and any multi-variant cis chromosome → "Other" (uncertain cis-combination). Diplotypes were formed by combining the two phased chromosomes of each individual, and CPIC-consistent metabolizer phenotypes (ultrarapid, rapid, normal, intermediate, poor) were assigned from the diplotype [4]. Frequencies were tabulated globally, per super-population, and per sub-population (32 groups in total).

Table 1. Haplotype encoding and functional definitions used to translate phased chromosomes into CYP2C19 star alleles. Bit order is `rs4244285`(`*2`) | `rs4986893`(`*3`) | `rs12248560`(`*17`).

Binary	Star allele	Function	Tag SNP (alt allele)
`000`	`*1`	Normal function	— (reference on all three)
`100`	`*2`	No function	`rs4244285` (A)
`010`	`*3`	No function	`rs4986893` (A)
`001`	`*17`	Increased function	`rs12248560` (T)
other	Other	Uncertain (cis)	≥ 2 alt alleles in cis

2.4 Missingness and Hardy–Weinberg equilibrium audit

Genotype and allele call-completeness were computed for every locus in every population group. HWE was tested for each of the three SNPs across all 32 strata (96 tests). Because two of the

three variants are rare or near-monomorphic in several populations, we used the exact HWE test of Wigginton, Cutler, and Abecasis rather than relying on a chi-square approximation, and additionally reported a Yates-corrected chi-square where expected cell counts were low (< 5) [23]. To control the family-wise error rate across the 96 tests we applied the Benjamini–Hochberg false-discovery-rate (FDR) procedure at $\alpha = 0.05$ [24]. Apparent HWE deviations were interpreted in the light of known population structure, since pooling divergent sub-populations can inflate global homozygosity through the Wahlund effect even when each sub-population is individually in equilibrium.

2.5 Tag-SNP fidelity and linkage-disequilibrium analysis

We quantified how faithfully each SNP tags its designated star allele by treating star-allele assignment on each of the 5,008 phased chromosomes as ground truth and computing per-SNP sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Pairwise linkage disequilibrium between the three tag SNPs was computed globally and within each super-population using the coefficient D, the normalized D' of Lewontin, and r^2 [25–27]. Standard population-genetic and association tooling conventions were followed for LD estimation on phased data [28].

2.6 Coverage estimation and uncertainty

Two complementary coverage metrics were defined. *Clinical carrier coverage* is the proportion of individuals in a group whose diplotype differs from `*1/*1` (i.e., who carry at least one star allele detectable by the core). *Haplotype-mass coverage* is the summed frequency of core-detectable star-allele haplotypes (`*2` + `*3` + `*17`) in the group. For carrier coverage, 95% confidence intervals were computed with the Wilson score method, which is preferred over the Wald interval for proportions near 0 or 1 and for small samples such as individual sub-populations [29, 30]. Total (global) coverage was then compared against worst-population coverage at both the super- and sub-population level.

2.7 Equity audit of uncovered loss-of-function alleles

To quantify the blind spot of the core panel, we assembled a candidate add-on set of no-function alleles—`*4` (`rs28399504`), `*6`, and `*8`—and validated each identifier against dbSNP before use [14, 22]. Two of the three task-provided identifiers were erroneous and were corrected to canonical PharmVar identifiers (Section 3.2). For every individual called `*1/*1` (normal metabolizer) by the core, we then asked whether that person nonetheless carried a genotypable uncovered no-function allele; such individuals constitute the *diagnostic gap* (false-normal misclassification). Uncovered-allele carrier rates and misclassification rates were tabulated globally, by super-population, and by sub-population, with Wilson confidence intervals. This audit was framed explicitly around ancestral equity, comparing the global blind spot against its magnitude in the African super-population and other worst-affected groups [11, 15].

2.8 Consistency verification and reproducibility

The final specification was assembled programmatically from the five upstream analysis summaries and subjected to 17 automated cross-file consistency checks (e.g., that each core-allele frequency matched between the haplotype and LD analyses, that liftover offsets were invariant, and that add-on-allele frequencies agreed between the coverage and equity modules). All 17 checks passed. Analyses were implemented in Python using standard scientific-computing libraries; every intermediate result was written to disk with full provenance.

3 Assay Specification

3.1 Core panel: universal *2/*3/*17

The recommended first-line panel comprises the three tag SNPs in Table 2. Each is a single, well-characterized biallelic variant with an unambiguous functional interpretation, dbSNP-verified coordinates on both reference builds, and a near-perfect tagging relationship to its star allele (Section 4.3). Together they capture the two most common no-function alleles and the common increased-function allele, and thus reconstruct the great majority of clinically relevant CYP2C19 diplotypes at minimal genotyping cost.

Table 2. Core CYP2C19 assay specification. Coordinates are dbSNP Build 157 authoritative positions; all three task-provided GRCh37 coordinates were incorrect and were corrected prior to extraction. Global allele frequency is estimated from 5,008 phased chromosomes of 1000 Genomes Phase 3.

Allele	rsID	GRCh37	GRCh38	Global freq.	Functional impact
*2	rs4244285	chr10:96,541,616 G>A	chr10:94,781,859 G>A	22.12%	No function; c.681G>A splice defect (exon 5)
*3	rs4986893	chr10:96,540,410 G>A	chr10:94,780,653 G>A	1.42%	No function; c.636G>A p.Trp212Ter (exon 4)
*17	rs12248560	chr10:96,521,657 C>T	chr10:94,761,900 C>T	15.30%	Increased function; c.-806C>T promoter

3.2 Population-specific add-on: *4/*6/*8

The trigger-activated add-on module targets three rare no-function alleles that the core cannot see (Table 3). Assembling this module surfaced two important identifier errors that any implementer must avoid. The task-provided identifier for *6 (rs142108216) maps to chromosome 4, not to the CYP2C19 locus on chromosome 10, and was rejected; its canonical PharmVar counterpart, rs72552267 (chr10:96,535,210 G>A), is nonetheless *absent* from the Phase 3 call set and therefore cannot be genotyped in this cohort at all. The task-provided identifier for *8 (rs11568732) is a common promoter variant (alternate-allele frequency ≈ 8%), biologically implausible for a rare loss-of-function allele, and was corrected to the canonical rs41291556 (chr10:96,535,173 T>C, frequency ≈ 0.1%). Only *4 and *8 are assessable in 1000 Genomes Phase 3; *6 is structurally unassessable in this reference, a fact with direct implications for the technology-escalation trigger (Section 6).

Table 3. Population-specific add-on specification. "Assessable" indicates whether the canonical variant is present in the 1000 Genomes Phase 3 call set and can therefore be genotyped/audited in this cohort.

Allele	Task rsID	Canonical rsID	GRCh37	Freq.	Assess.?	Functional impact
*4	rs28399504	rs28399504	chr10:96,522,463 A>G	0.080%	Yes	No function; c.1A>G initiation codon
*6	rs142108216	rs72552267	chr10:96,535,210 G>A	n/a	**No**	No function; c.395G>A p.Arg132Gln
*8	rs11568732	rs41291556	chr10:96,535,173 T>C	0.100%	Yes	No function; c.358T>C p.Trp120Arg

Implementer caution. Star-allele definitions are curated by rsID and genomic coordinate, not by trivial name. Two of the three add-on identifiers supplied to this analysis were wrong—one pointing to the wrong chromosome, one to a common variant of the wrong functional class. Any laboratory building this panel must validate every probe against the

current PharmVar/dbSNP records rather than trusting an inherited identifier list.

4 Results

4.1 Quality control: complete calls and no HWE deviation

The Phase 3 call set was of exemplary quality at these loci. Genotype and allele missingness were **zero** across all 2,504 individuals and all 32 population strata. Across the 96 HWE tests (3 SNPs × 32 groups), **no test** remained significant after Benjamini–Hochberg FDR correction at $\alpha = 0.05$ (Figure 1). The smallest raw p-values arose at the global level for `rs4244285` ($p = 0.014$) and `rs12248560` ($p = 0.019$), exactly where the Wahlund effect is expected to manifest most strongly because the global sample pools five divergent super-populations; both were rendered non-significant after correction (FDR $p = 0.56$). The rare `*3` tag `rs4986893`, near-monomorphic outside East Asia, triggered the exact-test path in 32 groups and showed no deviation anywhere. The joint observation of zero missingness and zero HWE deviation is strong evidence that the genotypes are free of the systematic artifacts—batch effects, mapping errors, or sample mix-ups—that would otherwise confound a coverage analysis.

4.2 Allele and haplotype frequencies vary strongly by ancestry

Global star-allele frequencies were `*1` 61.1%, `*2` 22.1%, `*17` 15.3%, and `*3` 1.4%, with a single "Other" cis chromosome (0.02%). These global figures, however, average over striking ancestral structure (Figure 2; Table 4). The no-function `*2` allele is roughly twice as common in South and East Asians (35.8% and 31.3%) as in Europeans (14.5%) or admixed Americans (10.4%). The increased-function `*17` allele shows the opposite gradient, peaking in Africans (23.5%) and Europeans (22.4%) and nearly vanishing in East Asians (1.5%). The `*3` allele is essentially an East-Asian-specific variant (5.6% in EAS) and is absent from the AMR and EUR super-populations in this cohort. This mirror-image distribution of the two functionally opposite common alleles is the central reason that core-assay yield is so ancestry-dependent.

Table 4. Star-allele frequencies by super-population (phased haplotypes). "Other" denotes the single uncertain cis-combination chromosome. Percentages are of chromosomes within each group.

Super-pop.	*1	*2	*3	*17	Other
AFR	59.23%	17.02%	0.23%	23.52%	0.00%
AMR	77.67%	10.37%	0.00%	11.82%	0.14%
EAS	61.71%	31.25%	5.56%	1.49%	0.00%
EUR	63.12%	14.51%	0.00%	22.37%	0.00%
SAS	49.39%	35.79%	1.23%	13.60%	0.00%
Global	**61.14%**	**22.12%**	**1.42%**	**15.30%**	**0.02%**

The sub-population heatmap (Figure 3) refines this picture and exposes the extremes that a coverage analysis must confront: the Peruvian (PEL) sample carries the highest `*1` (normal) frequency of any group (90.0% of haplotypes), while the Sri Lankan Tamil (STU) sample carries the lowest (43.1%). Translating diplotypes into predicted metabolizer phenotypes (Figure 4) shows the downstream consequence: the intermediate/poor-metabolizer burden is concentrated in East and South Asians (driven by `*2`/`*3`), whereas the rapid/ultrarapid burden is concentrated in Africans and Europeans (driven by `*17`). We report these phenotype distributions strictly as population-genetic summaries; *they carry no implication for any individual's prescribing.*

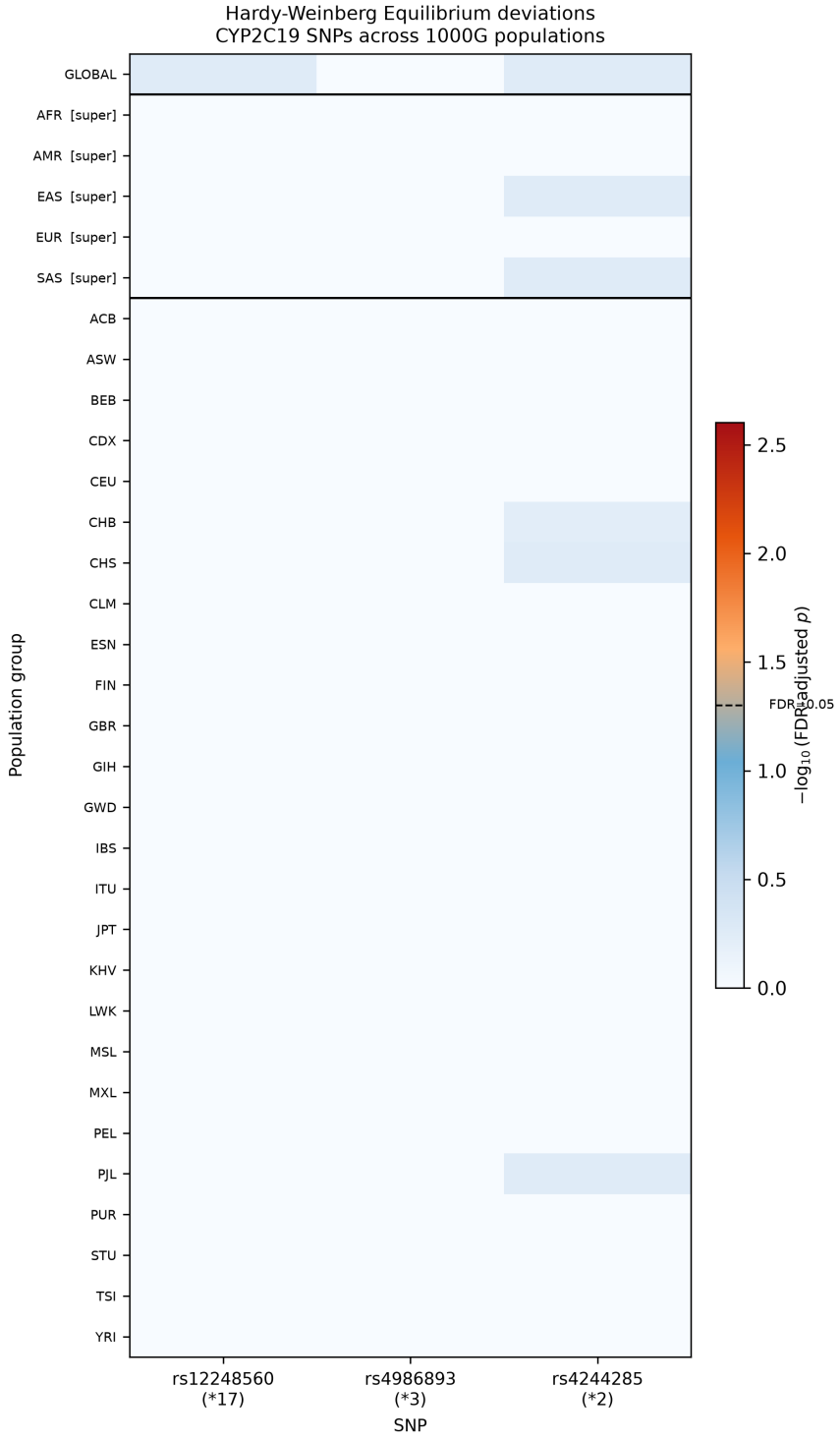


Figure 1. Hardy–Weinberg equilibrium audit across the three CYP2C19 tag SNPs (columns) and all 32 population groups (rows). Cell shading encodes $-\log_{10}$ of the FDR-adjusted p-value; the dashed reference marks the $\alpha = 0.05$ significance threshold. No cell crosses the threshold: there are zero significant HWE deviations after multiple-testing correction, consistent with high-quality, artifact-free genotype calls.

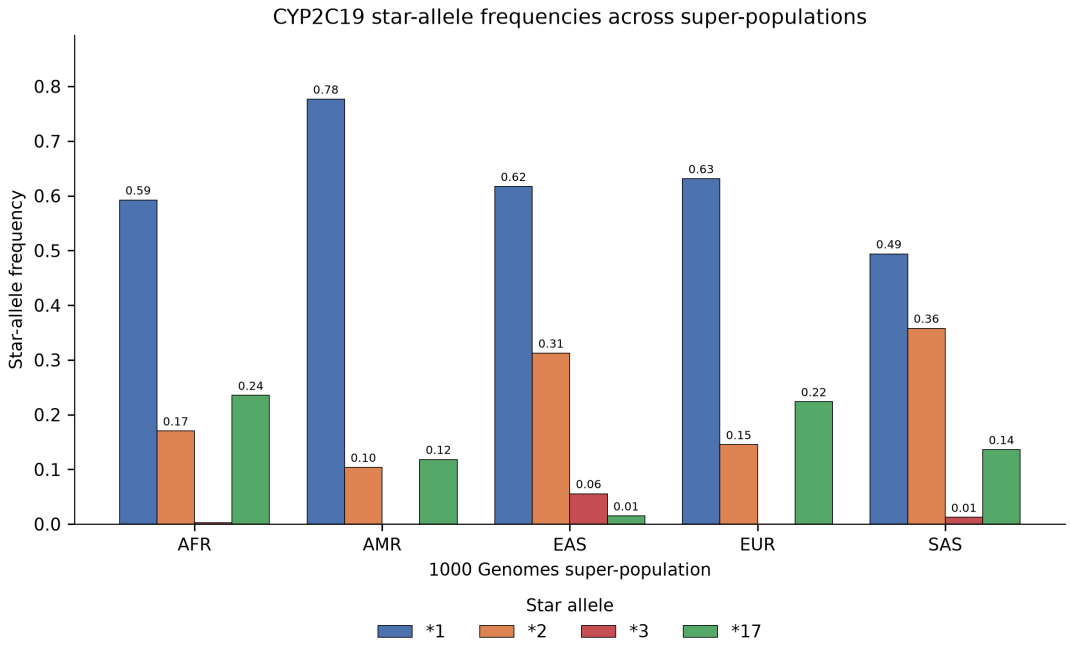


Figure 2. CYP2C19 star-allele frequencies across the five 1000 Genomes super-populations. The reciprocal gradients of the no-function `*2` allele (high in EAS/SAS) and the increased-function `*17` allele (high in AFR/EUR), together with the East-Asian concentration of `*3`, drive the population dependence of core-assay coverage.

4.3 Each SNP is a near-perfect tag for its star allele

Tag-SNP fidelity was essentially ideal (Table 5). Across all 5,008 phased chromosomes, every SNP tagged its star allele with sensitivity 1.000, and PPV was 0.9991 for `*2`, 1.0000 for `*3`, and 0.9987 for `*17`. The only imperfection in the entire dataset is a single chromosome (sample HG01198, PUR/AMR) carrying both the `*2` and `*17` alternate alleles in cis—the `101` haplotype—which is classified as "Other" rather than as either single star allele, and therefore contributes exactly one false positive to each of `*2` and `*17`. Excluding that one cis chromosome, PPV is 1.000 for all three SNPs. In practical terms, the three-SNP core is a faithful molecular readout of its intended star alleles; its limitation is not tagging error but the alleles it does not interrogate.

Table 5. Global tag-SNP fidelity over 5,008 phased chromosomes. TP/FP/TN/FN are true/false positives/negatives for the SNP-to-star-allele call. The lone false positive for `*2` and `*17` is the single `101` cis chromosome (HG01198) classified as "Other."

SNP	Allele	TP	FP	TN	FN	Sensitivity	PPV
`rs4244285`	`*2`	1108	1	3899	0	1.0000	0.9991
`rs4986893`	`*3`	71	0	4937	0	1.0000	1.0000
`rs12248560`	`*17`	766	1	4241	0	1.0000	0.9987

4.4 Linkage disequilibrium: the tag SNPs are mutually repulsive

The pairwise LD structure explains *why* three simple tag SNPs suffice (Figure 5). The defining alleles of `*2` and `*17` exhibit near-complete repulsion: globally $D' = -0.994$, and only that single `101` chromosome places the two alternate alleles in cis (out of 5,008). The `*3` tag likewise shows $D' = -1.0$ with both other alleles wherever it is polymorphic. This strong negative disequilibrium—D' at or near -1 throughout—is the population-genetic signature of mutually

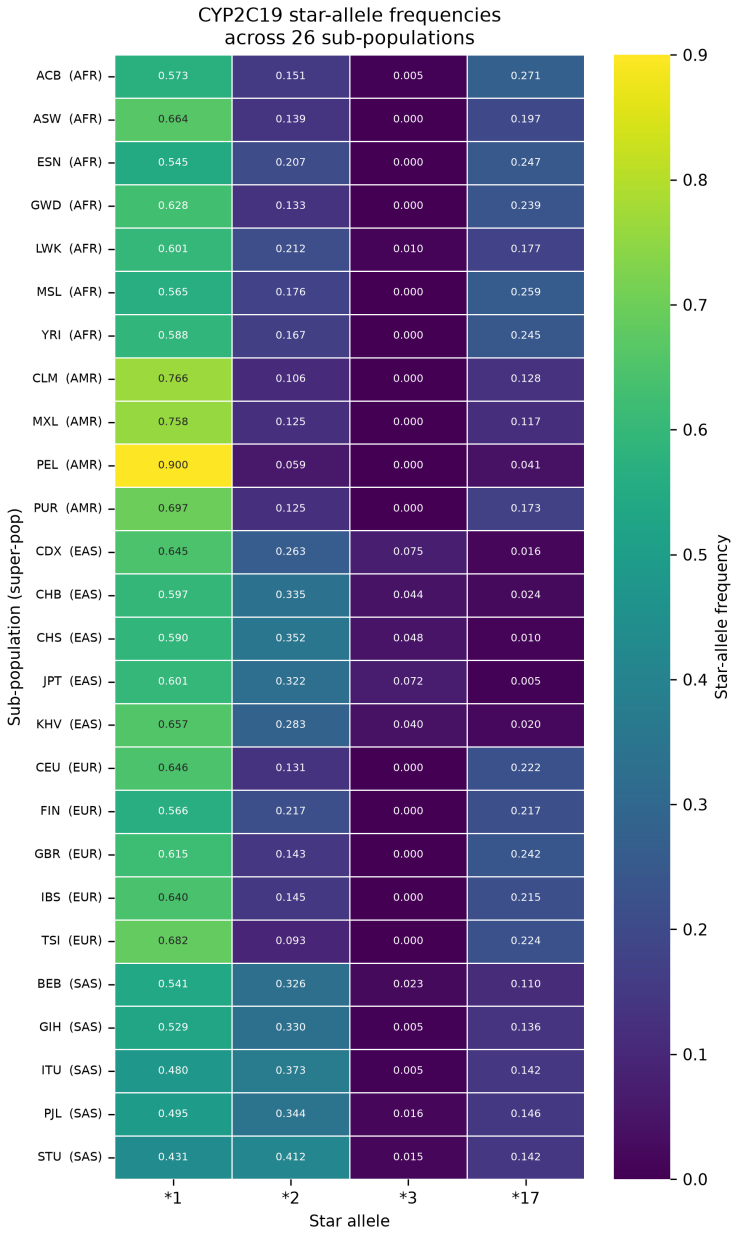


Figure 3. Star-allele frequencies across all 26 sub-populations (rows grouped by super-population). The normal-function `*1` column spans from 90.0% (PEL, AMR) to 43.1% (STU, SAS); because the core assay only detects non-`*1` alleles, this `*1` gradient is inversely equivalent to the core carrier-coverage gradient in Figure 6.

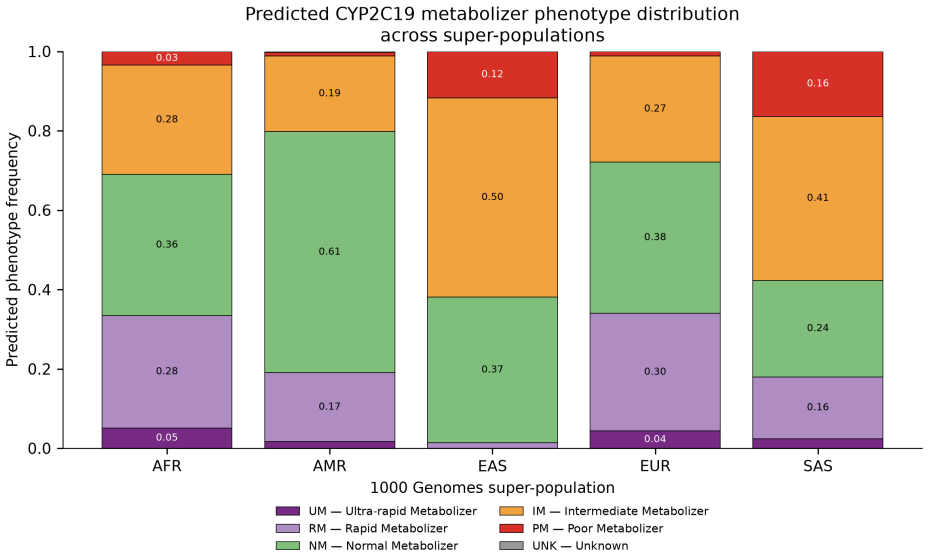


Figure 4. Predicted CYP2C19 metabolizer-phenotype distributions by super-population, derived from core-assay diplotypes under CPIC functional assignments. Shown as a descriptive population summary only. The reciprocal ancestral gradients in intermediate/poor versus rapid/ultrarapid categories follow directly from the `*2`/`*3` and `*17` frequency patterns.

exclusive haplotypes: each tag SNP marks its own haplotype background almost perfectly, so the three SNPs jointly enumerate distinct star alleles without ambiguity. The correspondingly low r^2 values (< 0.09 in every stratum) reflect the very different allele frequencies of the three variants and confirm that none of the SNPs is a redundant proxy for another; each contributes independent information. Where an allele is monomorphic in a super-population (e.g., `*3` in AMR and EUR), LD is undefined, as expected.

4.5 Coverage: a strong global average hiding a threefold ancestral spread

Globally, the core assay achieves a carrier coverage of **62.4%** (95% Wilson CI: 60.5–64.3%) and a haplotype-mass coverage of 38.8%. That headline number is respectable, but the per-population forest plot (Figure 6) reveals that it is an average over a nearly threefold range. At the super-population level, coverage runs from 75.7% in SAS down to **39.2%** in AMR (95% CI: 34.2–44.4%). At the sub-population level the spread is far wider: the best-covered group, STU (SAS), reaches 82.4%, while the worst-covered group, **PEL (AMR), reaches only 18.8%** (95% CI: 11.9–28.4%). The absolute gap between global and worst sub-population coverage is 43.60 percentage points, and the gap between best and worst sub-population is 63.50 points. Table 6 lists carrier coverage with Wilson intervals for the global cohort, all five super-populations, and the four worst- and three best-covered sub-populations.

The mechanism is transparent from the allele frequencies. Coverage is simply one minus the `*1`/`*1` frequency; because AMR populations—and PEL most acutely—carry very high `*1` (normal) frequencies and correspondingly little `*2`/`*3`/`*17`, a three-SNP panel keyed to those alleles finds little to report. In such a cohort, a negative core result poorly excludes atypical metabolism, because the actionable variation, to the extent it exists, is more likely to reside in rare or population-private alleles the panel does not carry.

5 Equity Audit

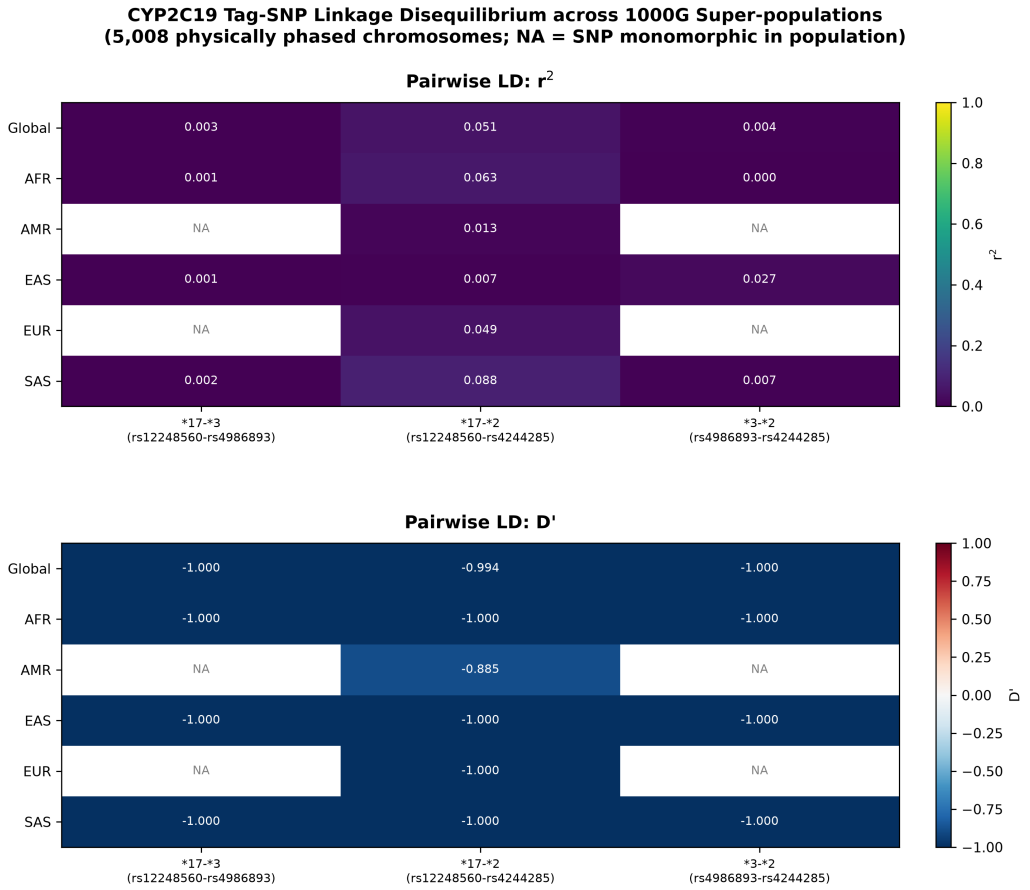


Figure 5. Pairwise linkage disequilibrium between the three tag SNPs, computed on 5,008 phased chromosomes globally and within each super-population. Top: r^2 (all values low, indicating non-redundant markers). Bottom: D' (at or near -1 throughout, indicating near-complete repulsion / mutually exclusive haplotypes). "NA" marks strata where a SNP is monomorphic.

Table 6. Core-assay carrier coverage (proportion of individuals with ≥ 1 detectable star allele) with 95% Wilson confidence intervals, for the global cohort, the five super-populations, and the four lowest- and three highest-covered sub-populations. Gap is the absolute difference from the global estimate.

Group	Super-pop.	n	Coverage	95% CI	Gap vs. global
Global (ALL)	—	2504	**62.4%**	[60.5, 64.3]	—
Super-populations					
SAS	SAS	489	75.7%	[71.7, 79.3]	+13.3
AFR	AFR	661	64.4%	[60.7, 68.0]	+2.1
EAS	EAS	504	63.3%	[59.0, 67.4]	+0.9
EUR	EUR	503	61.8%	[57.5, 66.0]	−0.6
AMR	AMR	347	39.2%	[34.2, 44.4]	−23.2
Lowest-covered sub-populations					
PEL	AMR	85	**18.8%**	[11.9, 28.4]	**−43.6**
MXL	AMR	64	40.6%	[29.5, 52.9]	−21.8
CLM	AMR	94	41.5%	[32.1, 51.6]	−20.9
PUR	AMR	104	52.9%	[43.4, 62.2]	−9.5
Highest-covered sub-populations					
ITU	SAS	102	78.4%	[69.5, 85.3]	+16.0
PJL	SAS	96	75.0%	[65.5, 82.6]	+12.6
STU	SAS	102	**82.4%**	[73.8, 88.5]	+20.0

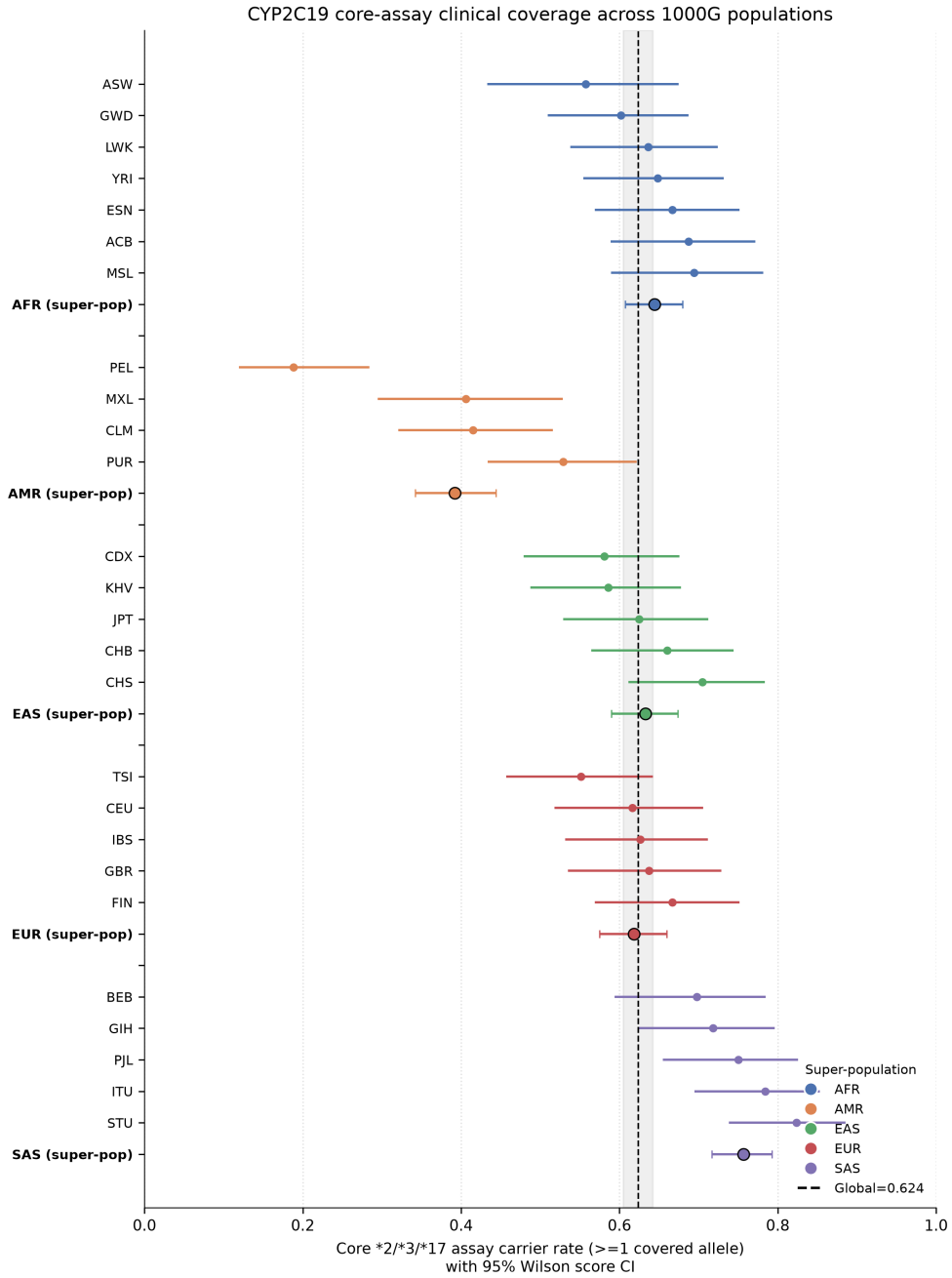


Figure 6. Core-assay carrier coverage across 1000 Genomes populations with 95% Wilson confidence intervals; sub-populations are grouped by super-population and super-population summaries are shown in bold. The dashed line marks the global estimate (62.4%). The Peruvian (PEL) sample at 18.8% is a striking negative outlier, and the entire AMR super-population sits well below the global average, whereas South Asian groups sit above it.

5.1 The core is blind to a defined set of loss-of-function alleles

The coverage analysis measures what the core *detects*; the equity audit measures what it *misses*, and for whom. Restricting to the two uncovered no-function alleles that are genotypable in this cohort (`*4` at 0.080% and `*8` at 0.100% global frequency), we identified **eight individuals** worldwide who are called `*1/*1` (normal metabolizer) by the core yet carry a genuine uncovered no-function allele. These eight constitute a *diagnostic gap*: a false-normal misclassification rate of 0.319% across the whole cohort, or 0.849% among the subset the core labels normal. While small in absolute terms, this residual gap is precisely the population for whom a low-cost panel silently returns a falsely reassuring result.

5.2 The measured gap is a floor, not a ceiling—and it is ancestrally structured

Two features of the audit matter more than the raw count. First, the distribution of the uncovered burden is ancestrally uneven (Figure 7). The uncovered no-function alleles that *are* genotypable here happen to appear across several super-populations (with, in this particular cohort, the highest per-individual misclassification observed in a European sub-population), but this is an artifact of which alleles the reference panel can see. Second, and decisively, the canonical `*6` allele (`rs72552267`) is **entirely absent** from the 1000 Genomes Phase 3 call set. Because `*6` and other rare no-function alleles are known to occur at appreciable frequency in African and other under-represented populations, and because array- and imputation-based reference data cannot detect a variant that is not in the reference at all, **the true diagnostic gap for these populations is systematically underestimated by any analysis built on this reference** [11, 14, 15]. The equity concern is therefore not merely the eight individuals we can count; it is the unknown, structurally invisible burden we cannot.

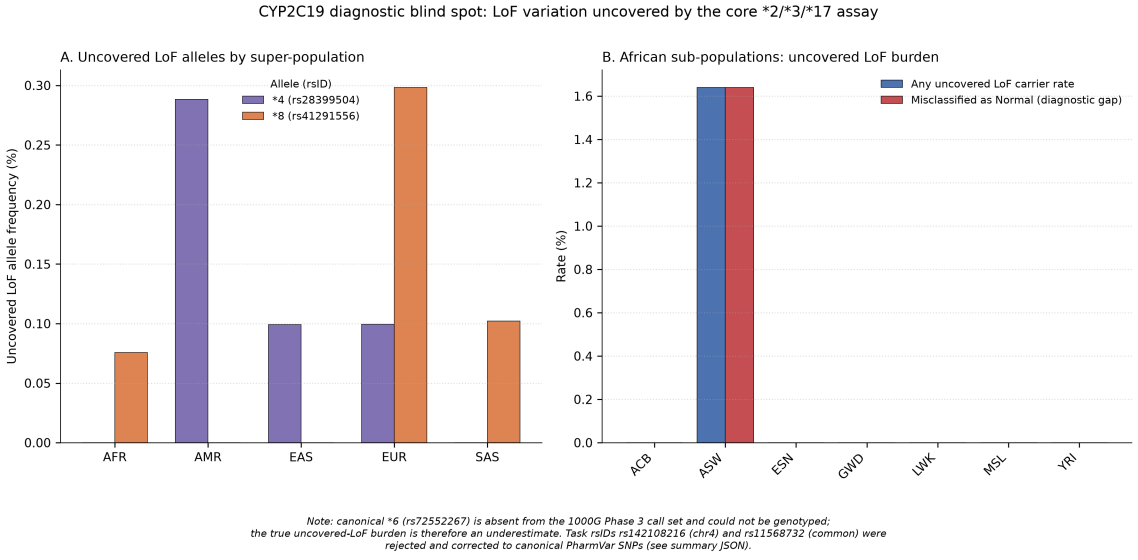


Figure 7. Diagnostic blind spot of the core `*2/*3/*17` assay. (A) Frequencies of the two genotypable uncovered no-function alleles (`*4`, `*8`) by super-population. (B) Uncovered loss-of-function carrier and misclassification rates across African sub-populations. The canonical `*6` allele (`rs72552267`) is absent from the Phase 3 call set and cannot be genotyped, so the true uncovered burden—especially for African ancestry—is underestimated here.

5.3 Coverage inequity is the dominant equity signal

Even setting aside the rare-allele blind spot, the coverage analysis of Section 4.5 is itself an equity finding. A universal-only panel would deliver 82% clinical yield to a Sri Lankan Tamil research cohort but only 19% to a Peruvian one. Framed as a research instrument, the core assay is not equally informative across ancestries: for admixed American and especially Indigenous-American-ancestry cohorts, a negative result carries far less information, and the panel's design implicitly privileges the ancestries in which `*2`, `*3`, and `*17` happen to be common. Recognizing and quantifying this differential yield—rather than reporting only the flattering 62.4% global average—is the core ethical obligation of an equity-aware assay design [1, 12, 13].

Equity bottom line. The core assay's weakness is not technical error—it tags its alleles almost perfectly—but *differential coverage*. Its clinical yield ranges from 18.8% (PEL) to 82.4% (STU), and its blind spot toward rare, often African-enriched no-function alleles (notably the unassessable `*6`) is larger than any reference-based count can show. A design that reports only the global average would mask both problems.

6 Decision Framework: The Evidence That Would Change the Design

The analysis supports a layered design rather than a binary choice. The universal core is the correct, cost-effective baseline everywhere; defined, quantitative triggers then determine when a study or program should add the `*4/*6/*8` module or move to whole-gene sequencing. We organize the triggers along three axes—ancestry, study utility, and technology—and summarize them in Table 7 and the decision tree of Figure/Box below.

6.1 Baseline recommendation

Deploy the three-SNP core (`*2`, `*3`, `*17`) as the universal first-line panel. It captures 38.8% of star-allele haplotype mass, identifies a variant in 62.4% of individuals globally (95% CI: 60.5–64.3%), and tags each allele with sensitivity ≈ 1.0 and PPV ≥ 0.999. For the majority of well-characterized, ancestrally European, East Asian, or South Asian research cohorts, the core alone is defensible provided its coverage and limitations are documented.

6.2 Ancestry-based triggers

Two ancestry conditions should change the design. When a target cohort contains more than roughly 5% African ancestry, the `*4/*6/*8` add-on should be *mandated*, because African populations carry rare no-function alleles—notably `*6` and `*8`—that the core cannot see, and omitting them systematically under-ascertains poor-metabolizer status in individuals of African descent. Critically, because `*6` is unassessable in the Phase 3 reference, this trigger's quantitative anchor is a lower bound and the add-on should be paired with sequencing (technology axis) wherever feasible. Separately, when a cohort has high Indigenous-American ancestry (PEL-like AMR profiles), the analysis should *flag the core's low utility*: PEL coverage is only 18.8% and the AMR super-population averages 39.2%, so a near-uniformly negative core poorly excludes atypical metabolism and whole-gene sequencing (or at minimum an expanded panel) is preferred.

6.3 Utility-based triggers

Two study-design conditions should also change the design. If the *projected* core carrier rate in the target cohort falls below ∼40% (the AMR super-population level), the incremental yield of additional and rare variants is proportionally high, so the add-on should be engaged; if

the projected rate falls below 30%, whole-gene sequencing is warranted. These thresholds are anchored on the observed extremes (PEL 18.8% to STU 82.4%). Independently, whenever the study endpoint requires confidently *ruling out* atypical metabolism—for example, research on antiplatelet, proton-pump-inhibitor, or antidepressant response where a false-normal call carries analytic risk—the add-on should be included regardless of ancestry, and the residual diagnostic-gap rate (0.849% among core-normal individuals) reported as an explicit limitation [5, 6, 31, 32].

6.4 Technology-based triggers

Finally, two technology conditions favor abandoning the fixed panel altogether. When per-sample high-throughput sequencing or targeted-capture cost falls to approximate parity with multiplex genotyping, the program should transition to gene-level next-generation or long-read sequencing of the CYP2C19 locus, because only sequencing can capture `*6` and other variants that fixed panels and imputation references structurally miss—the absence of `rs72552267` from Phase 3 being the concrete example. And for any pan-ancestry or discovery cohort in which private or novel variants are expected—especially understudied populations—whole-gene sequencing should be the default, since a fixed panel encodes prior assumptions about which variants matter that fail exactly where reference data are sparsest [15, 17].

Table 7. Decision-framework triggers for escalating beyond the universal core assay. Each trigger is anchored to a quantitative observation from this analysis.

ID	Axis	Trigger	Action & quantitative anchor
ANC-1	Ancestry	$> 5\%$ African ancestry in cohort	Mandate `*4/*6/*8` add-on; pair with sequencing (`*6` unassessable).
ANC-2	Ancestry	High Indigenous-American / PEL-like AMR ancestry	Flag low utility; recommend whole-gene sequencing. PEL coverage 18.8%; AMR 39.2%.
UTIL-1	Utility	Projected core carrier rate $< 40\%$	Add-on panel; if $< 30\%$, escalate to sequencing. Observed range 18.8–82.4%.
UTIL-2	Utility	Endpoint requires confident exclusion of atypical metabolism	Include add-on regardless of ancestry; report residual gap (0.849% of core-normal).
TECH-1	Technology	Sequencing cost $\approx$ genotyping cost	Transition to gene-level NGS/long-read; captures `*6` & private variants.
TECH-2	Technology	Pan-ancestry / discovery / understudied cohort	Default to whole-gene sequencing rather than any fixed panel.

Decision tree.

1. **Start:** deploy the universal core `*2/*3/*17` assay.
2. African ancestry $> 5\%$? $\rightarrow$ add `*4/*6/*8` module (ANC-1).
3. High Indigenous-American/AMR ancestry, or projected core coverage $< 40\%$? $\rightarrow$ flag low utility; add-on, and if $< 30\%$ escalate to whole-gene sequencing (ANC-2 / UTIL-1).
4. Endpoint requires confident exclusion of atypical metabolism? $\rightarrow$ add module and report residual gap (UTIL-2).
5. Sequencing cost at parity, or discovery/understudied cohort? $\rightarrow$ transition to gene-level sequencing to capture `*6` and private variants (TECH-1 / TECH-2).
6. Otherwise $\rightarrow$ the core is sufficient; document coverage and limitations.

7 Discussion

7.1 Principal findings

This analysis resolves the design question with nuance: a universal *2/*3/*17 core is the right *baseline* but not a sufficient *universal* answer. The core is inexpensive and technically excellent—zero missingness, no HWE deviation, and near-perfect tagging with D' near -1 confirming that three SNPs cleanly enumerate mutually exclusive haplotypes. Its limitation is entirely one of content coverage, and that limitation is ancestrally patterned. A single global coverage figure of 62.4% is, by itself, a misleading summary; the honest description is a distribution running from 18.8% to 82.4%, with the admixed-American super-population and Indigenous-American-ancestry sub-populations systematically disadvantaged, and with a rare-allele blind spot—epitomized by the unassessable *6—that falls hardest on the very populations least represented in reference data.

7.2 Relationship to prior work

Our super-population allele frequencies are concordant with large worldwide surveys of CYP2C19 variation, including the enrichment of *2/*3 in East and South Asian ancestry and of *17 in African and European ancestry [14, 16]. The observation that rare, population-specific alleles contribute materially to functional variation, and that they are enriched in under-represented populations, matches recent population-scale sequencing analyses of the pharmacogene landscape [15]. More broadly, the equity pattern we quantify for a single pharmacogene is a concrete instance of the field-wide diversity gap that has been documented for GWAS, polygenic scores, and clinical genomics [11–13]. The functional interpretation of the star alleles rests on CPIC/PharmVar curation [3, 4, 8].

7.3 Limitations

Several limitations bound the interpretation. First, and most importantly, the reference cohort constrains what can be measured: the canonical *6 variant is absent from 1000 Genomes Phase 3, so every uncovered-burden figure here is a lower bound, and the true African diagnostic gap is larger than reported. Second, 1000 Genomes sub-population samples are modest in size (often 60–110 individuals), which widens the Wilson intervals for the extreme groups—PEL's coverage interval spans 11.9–28.4%—and cautions against over-precise point estimates. Third, the phenotype distributions we present are population-genetic summaries under CPIC functional rules and physical phasing; they are not, and must not be read as, individual metabolizer predictions. Fourth, the analysis considers only the specified tag SNPs and add-on alleles; the full PharmVar CYP2C19 star-allele catalogue is larger, and additional rare alleles would further reduce core coverage in the direction of a stronger case for sequencing. Finally, the identifier errors we encountered—one add-on rsID on the wrong chromosome, one of the wrong functional class—underscore that even the inputs to such analyses require independent verification.

7.4 Implications for assay design

The practical implication is a tiered, ancestry-aware deployment. Fund and field the cheap core universally, but (i) never report the global average without the per-population distribution and its uncertainty; (ii) attach the *4/*6/*8 module wherever African ancestry is non-trivial or the endpoint demands confident exclusion of atypical metabolism; (iii) treat low projected coverage (AMR-like cohorts) as an explicit signal that the core is under-informative; and (iv) plan an orderly transition to gene-level sequencing as its cost approaches that of genotyping, since sequencing is the only design that dissolves the reference-panel blind spot rather than merely documenting it [1, 2].

8 Conclusion

A low-cost, research-only CYP2C19 assay should be built on a universal `*2/*3/*17` core, augmented by a trigger-activated `*4/*6/*8` add-on rather than deployed as a single fixed universal panel. The core is cheap and technically near-perfect, but its clinical yield ranges from 18.8% to 82.4% across populations, and its blind spot toward rare no-function alleles—including the reference-invisible `*6`—is ancestrally structured and underestimated by array-based data. Equitable design therefore requires reporting per-population coverage with uncertainty, activating population-specific content where ancestry or study endpoint demands it, and transitioning to whole-gene sequencing as cost permits. The evidence that would change the design is explicit and quantitative: African ancestry above 5%, projected coverage below 40% (add-on) or 30% (sequencing), endpoints requiring confident exclusion of atypical metabolism, and sequencing reaching cost parity with genotyping. **This report is a research and design analysis and provides no prescribing advice.**

Data, Provenance, and Reproducibility

All analyses derive from the public 1000 Genomes Project Phase 3 chromosome-10 release (20130502, GRCh37) and its sample panel, with variant annotation cross-checked against dbSNP Build 157 and functional interpretation against CPIC/ClinPGx record `PA166251443`. The workflow proceeded in six documented stages—data acquisition and coordinate verification, missingness/HWE audit, phased haplotype and diplotype estimation, tag-SNP fidelity and LD analysis, coverage and equity audit, and specification consolidation—and the consolidated specification passed all 17 automated cross-file consistency checks. Intermediate tables (per-population coverage, HWE results, LD matrices, tagging metrics, diagnostic-gap counts) and all figures were written to disk to permit independent verification.

Non-Clinical Use Statement

This document is intended solely for research and assay-design purposes. It reports population-level allele frequencies, coverage, statistical uncertainty, and an equity audit for a candidate genotyping panel. It is not a validated diagnostic, does not constitute medical advice, and must not be used to guide prescribing, dosing, or clinical management of any individual. Metabolizer-phenotype distributions are descriptive population summaries only. Any clinical application of CYP2C19 genotyping must use an appropriately validated, regulatory-cleared assay interpreted by qualified professionals under current clinical guidelines.

References

[1] Dan M. Roden, Howard L. McLeod, Mary V. Relling, Marc S. Williams, George A. Mensah, Josh F. Peterson, and Sara L. Van Driest. Pharmacogenomics. *The Lancet*, 394(10197): 521–532, 2019. doi: 10.1016/S0140-6736(19)31276-0.

[2] Mary V. Relling and William E. Evans. Pharmacogenomics in the clinic. *Nature*, 526 (7573):343–350, 2015. doi: 10.1038/nature15817.

[3] Andrea Gaedigk, Magnus Ingelman-Sundberg, Neil A. Miller, J. Steven Leeder, Michelle Whirl-Carrillo, Teri E. Klein, and PharmVar Steering Committee. The pharmacogene variation (PharmVar) consortium: Incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clinical Pharmacology & Therapeutics*, 103(3):399–401, 2018. doi: 10.1002/cpt.910.

[4] Craig R. Lee, Jasmine A. Luzum, Katrin Sangkuhl, Roseann S. Gammal, Marc S. Sabatine, C. Michael Stein, David F. Kisor, Nita A. Limdi, Yee Ming Lee, Stuart A. Scott, Jean-Sébastien Hulot, Dan M. Roden, Andrea Gaedigk, Kelly E. Caudle, Teri E. Klein, Julie A. Johnson, and Alan R. Shuldiner. Clinical pharmacogenetics implementation consortium guideline for CYP2C19 genotype and clopidogrel therapy: 2022 update. *Clinical Pharmacology & Therapeutics*, 112(5):959–967, 2022. doi: 10.1002/cpt.2526.

[5] John J. Lima, Cameron D. Thomas, Julia Barbarino, Zeruesenay Desta, Sara L. Van Driest, Nihal El Rouby, Julie A. Johnson, Larisa H. Cavallari, Valentina Shakhnovich, David L. Thacker, Stuart A. Scott, Matthias Schwab, Chakradhara Rao S. Uppugunduri, Christine M. Formea, James P. Franciosi, Katrin Sangkuhl, Andrea Gaedigk, Teri E. Klein, Roseann S. Gammal, and Takahisa Furuta. Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2C19 and proton pump inhibitor dosing. *Clinical Pharmacology & Therapeutics*, 109(6):1417–1423, 2021. doi: 10.1002/cpt.2015.

[6] J. Kevin Hicks, Jesse J. Swen, Andrea Gaedigk, Henry M. Dunnenberger, Teri E. Klein, Sook Wah Lee, Maria L. Alvarellos, Andria L. Del Tredici, Todd C. Skaar, Daniel J. Müller, Mia Wadelius, Gilbert J. Burckart, and Julie A. Johnson. Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2D6 and CYP2C19 genotypes and dosing of selective serotonin reuptake inhibitors. *Clinical Pharmacology & Therapeutics*, 98(2): 127–134, 2015. doi: 10.1002/cpt.147.

[7] Chad A. Bousman, James M. Stevenson, Laura B. Ramsey, Katrin Sangkuhl, J. Kevin Hicks, Jeffrey R. Strawn, Ajeet B. Singh, Gualberto Ruaño, Jacob T. Brown, Gillian C. Bell, J. Steven Leeder, Andrea Gaedigk, Stuart A. Scott, Teri E. Klein, Kelly E. Caudle, and Jeffrey R. Bishop. Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2D6, CYP2C19, CYP2B6, SLC6A4, and HTR2A genotypes and serotonin reuptake inhibitor antidepressants. *Clinical Pharmacology & Therapeutics*, 114(1):51–68, 2023. doi: 10.1002/cpt.2903.

[8] Michelle Whirl-Carrillo, Rachel Huddart, Li Gong, Katrin Sangkuhl, Caroline F. Thorn, Ryan Whaley, and Teri E. Klein. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 110(3):563–572, 2021. doi: 10.1002/cpt.2350.

[9] Sergio M. F. de Morais, Grant R. Wilkinson, Joyce Blaisdell, Kazuko Nakamura, Urs A. Meyer, and Joyce A. Goldstein. The major genetic defect responsible for the polymorphism of S-mephenytoin metabolism in humans. *Journal of Biological Chemistry*, 269(22):15419–15422, 1994. doi: 10.1016/S0021-9258(17)40694-6.

[10] Sarah C. Sim, Christer Risinger, Marja-Liisa Dahl, Eleni Aklillu, Mona Christensen, Leif Bertilsson, and Magnus Ingelman-Sundberg. A common novel CYP2C19 gene variant causes ultrarapid drug metabolism relevant for the drug response to proton pump inhibitors and antidepressants. *Clinical Pharmacology & Therapeutics*, 79(1):103–113, 2006. doi: 10.1016/j.clpt.2005.10.002.

[11] Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019. doi: 10.1016/j.cell.2019.02.048.

[12] Alice B. Popejoy and Stephanie M. Fullerton. Genomics is failing on diversity. *Nature*, 538 (7624):161–164, 2016. doi: 10.1038/538161a.

[13] Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, 2019. doi: 10.1038/s41588-019-0379-x.

[14] Ingrid Fricke-Galindo, Carolina Céspedes-Garro, Fernanda Rodrigues-Soares, María-Eugénia G. Naranjo, Alejandro Delgado, Fernando de Andrés, Marisol López-López, Eva Peñas Lledó, and Adrián LLerena. Interethnic variation of CYP2C19 alleles, 'predicted' phenotypes and 'measured' metabolic phenotypes across world populations. *The Pharmacogenomics Journal*, 16(2):113–123, 2016. doi: 10.1038/tpj.2015.70.

[15] Yitian Zhou and Volker M. Lauschke. The genetic landscape of major drug metabolizing cytochrome P450 genes—an updated analysis of population-scale sequencing data. *The Pharmacogenomics Journal*, 22(5-6):284–293, 2022. doi: 10.1038/s41397-022-00288-2.

[16] Mohitosh Biswas and Md Siddiqul Karim Kali. Global distribution of CYP2C19 risk phenotypes affecting safety and effectiveness of medications. *The Pharmacogenomics Journal*, 21(2):190–199, 2021. doi: 10.1038/s41397-020-00196-3.

[17] The H3Africa Consortium. Enabling the genomic revolution in africa. *Science*, 344(6190): 1346–1348, 2014. doi: 10.1126/science.1251546.

[18] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi: 10.1038/nature15393.

[19] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012. doi: 10.1038/nature11632.

[20] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, 2012. doi: 10.1038/nmeth.1785.

[21] Petr Danecek, Adam Auton, Gonçalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27 (15):2156–2158, 2011. doi: 10.1093/bioinformatics/btr330.

[22] Stephen T. Sherry, Minghong H. Ward, Michael Kholodov, John Baker, Lon Phan, Elizabeth M. Smigielski, and Karl Sirotkin. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001. doi: 10.1093/nar/29.1.308.

[23] Janis E. Wigginton, David J. Cutler, and Gonçalo R. Abecasis. A note on exact tests of hardy-weinberg equilibrium. *American Journal of Human Genetics*, 76(5):887–893, 2005. doi: 10.1086/429864.

[24] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[25] Richard C. Lewontin. The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics*, 49(1):49–67, 1964. doi: 10.1093/genetics/49.1.49.

[26] Christopher S. Carlson, Michael A. Eberle, Mark J. Rieder, Qian Yi, Leonid Kruglyak, and Deborah A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, 74(1):106–120, 2004. doi: 10.1086/381000.

[27] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005. doi: 10.1038/nature04226.

[28] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, 2007. doi: 10.1086/519795.

[29] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[30] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001. doi: 10.1214/ss/1009213286.

[31] Jessica L. Mega, Sandra L. Close, Stephen D. Wiviott, Lei Shen, Richard D. Hockett, John T. Brandt, Joseph R. Walker, Elliott M. Antman, William Macias, Eugene Braunwald, and Marc S. Sabatine. Cytochrome P-450 polymorphisms and response to clopidogrel. *New England Journal of Medicine*, 360(4):354–362, 2009. doi: 10.1056/NEJMoa0809171.

[32] Stuart A. Scott, Katrin Sangkuhl, C. Michael Stein, Jean-Sébastien Hulot, Jessica L. Mega, Dan M. Roden, Teri E. Klein, Marc S. Sabatine, Julie A. Johnson, and Alan R. Shuldiner. Clinical pharmacogenetics implementation consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. *Clinical Pharmacology & Therapeutics*, 94(3):317–323, 2013. doi: 10.1038/clpt.2013.105.